

Social media mining for identification and exploration of health-related information from pregnant women

Pramod Chandrashekar
Department of Biomedical
Informatics
Arizona State University
Tempe, AZ, USA
pbchandr@asu.edu

Arjun Magge
Department of Biomedical
Informatics
Arizona State University
Tempe, AZ, USA
amagge@asu.edu

Abeed Sarker
Department of Biostatistics
and Epidemiology
University of Pennsylvania
Philadelphia, PA, USA
abeed@upenn.edu

Graciela Gonzalez
Department of Biostatistics
and Epidemiology
University of Pennsylvania
Philadelphia, PA, USA
gragon@upenn.edu

ABSTRACT

Widespread use of social media has led to the generation of substantial amounts of information about individuals, including health-related information. Thus, social media provides the opportunity to study health-related information about selected population groups who may be of interest for a particular study. In this paper, we explore the possibility of utilizing social media data to perform targeted data collection and analysis from a particular population group—pregnant women. We hypothesize that we can use social media to identify cohorts of pregnant women and follow them over time to analyze crucial health-related information. To identify potentially pregnant women, we employ simple rule-based searches that attempt to detect pregnancy announcements with moderate precision. To further filter out false positives and noise, we employ a supervised classifier using a small number of hand-annotated data. Following the identification of a reasonably sized cohort, we collect their posts over time to create longitudinal health timelines and attempt to divide the timelines into different pregnancy trimesters. Finally, we assess the usefulness of the timelines by performing a preliminary analysis to estimate drug intake patterns of our cohort at different trimesters. Our rule-based cohort identification technique collected 53,820 users over thirty months from Twitter. Our pregnancy announcement classification technique achieved an F-measure of 0.81 for the pregnancy class, resulting in 34,895 user timelines. Analysis of the timelines revealed that pertinent health-related information, such as drug-intake and adverse reactions can be mined from the data. Our approach to using user timelines in this fashion has produced very encouraging results and can be employed for an array of other important tasks where cohorts, for which health-related information may not be available from other sources, are required to be followed over time to derive population-based estimates.

1. INTRODUCTION

Pre-market clinical trials assess the safety of drugs/medications (we use the terms interchangeably in this paper) in limited settings, and so the effects of those drugs on particular

cohorts (*e.g.*, pregnant women, children, or people suffering from specific conditions) cannot be assessed. Spontaneous reporting systems, such as the FDA Adverse Event Reporting System (FAERS), are used for post-marketing drug safety surveillance and they provide a mechanism for reporting adverse events associated with medication consumption. Although these sources may accumulate drug safety knowledge about specific population groups, studies have shown that they suffer from various problems, such as under-reporting [15]. To overcome these problems, additional sources of information are being actively utilized for pharmacovigilance tasks. Studies have shown that 26% of online adults discuss health information using social media [7], with approximately 90% women using online media for health-care information, and 60% using pregnancy related apps for support. These statistics suggest that social media sources may contain key information regarding specific cohorts, such as pregnant women, and their drug usage habits.

Although consuming drugs during pregnancy is not recommended by doctors worldwide, their usage is commonplace. For example, during pregnancy, women continue taking prescription drugs for ailments which preceded the pregnancy. For common health problems like heartburn, common cold and body pains, women tend to take over-the-counter medicines which may cause harm to the fetus. Past research has also indicated that 50% of the pregnancies in the United States are unintended [11]. In such cases, the fetus may be exposed to drugs without the mother's explicit knowledge. Currently, the U.S FDA maintains a list of pregnancy exposure registries¹ that collect health information on exposure to medical products during pregnancy. Such registries require pregnant women to voluntarily sign up, and hence, they suffer from low enrollment and follow-up rates [35]. Considering the fact that infant mortality rates are estimated to be at 5.96 deaths per 1,000 live births [1], and that the causes of 50% of these birth defects are unknown [22], identifying and utilizing additional sources for monitoring health information of pregnant women, such as social

¹<http://www.fda.gov/ScienceResearch/SpecialTopics/WomensHealthResearch/ucm134848.htm>

media, is of paramount importance.

A very popular social network, that is currently being extensively used for public health monitoring tasks, is Twitter—a micro-blogging site which is actively used by over 313 million users.² Despite the noisy nature of data on Twitter, because of the high volume and frequency, it is an attractive resource for big data mining tasks. In addition to widely used social networks like Twitter, there are also *online health communities*, which facilitate health-related information sharing over the Internet. One such online health community is DailyStrength³, which has over 400,000 members engaging in discussions among its 500+ groups. In contrast to tweets, the posts in online health forums like DailyStrength have no strict constraints on word counts. The language used is more formal and the availability of domain-specific discussion forums increase the chances of finding relevant medical information from discussions [28]. Thus, Twitter and DailyStrength present quite different types of social media chatter, with the data from the latter being significantly lower in terms of both volume and noise. Both these data sources carry health-related knowledge expressed by various cohorts but require customized techniques for mining the knowledge encapsulated.

1.1 Motivation, Goals and Contributions

Given the limited amount of information that is available about pregnant women during pre-market clinical trials, there is a need to explore additional resources of information. The presence of large amounts of social media data, which hold crucial health-related information, presents a strong motivation for developing frameworks for mining longitudinal information from this resource. Based on these motivations, the goals of this paper are as follows:

- Develop natural language processing (NLP), machine learning, and information retrieval (IR) methods for accurately identifying a cohort of pregnant women and collecting their social media timelines.
- Perform preliminary analyses of the extracted health timelines to assess their usefulness, identify limitations, and establish future research goals.

The main contributions of the paper are as follows:

- We present a framework by which social media data can be used to identify and collect information about pregnant women.
- We show that health timelines collected from social media contain crucial health-related information, which may be used in longitudinal studies.
- We discuss techniques for further dividing the timelines into pregnancy trimesters and verify that trimester-specific information can also be mined from the timelines.
- We discuss the current limitations of our novel idea and outline future directions

²<https://about.twitter.com/company>. Accessed on: 03/03/2016.

³<http://www.dailystrength.org>. Accessed on: 03/03/2016.

The rest of the paper is organized as follows: in Section 2, we briefly outline past research related to ours, including social media mining and data-centric approaches to pregnancy-safety monitoring; in Section 3, we detail our methods for identifying cohorts from the two social media sources, extracting relevant longitudinal data, and analyzing the data in a preliminary fashion; in Section 4, we present our results and provide discussions regarding our plans to build on this pilot for larger future projects; in Section 5, we discuss the limitations of our work and outline some planned future work; and we conclude the paper in section 6.

2. RELATED WORK

Research work most closely related to ours is in the domain of pharmacovigilance from social media, although, to the best of our knowledge, no past research has attempted to identify and follow longitudinal cohort information from this domain. Most of the research in pharmacovigilance and drug safety surveillance has focused on identifying adverse reactions associated with medications. Some past research has attempted to employ classification techniques to determine adverse drug reaction (ADR) assertive posts. For these tasks, two primary techniques have been attempted: lexicon-based classification and supervised classification. In lexicon-based classifications [25, 19] a given text is classified as having an ADR if it meets a set of specified lexical rules. Supervised classification techniques [33, 6] involve training classifiers using features from annotated data (used as training data) to automatically make classification decisions on test data based on observed probabilities in the training data.

Due to the advances in natural language processing (NLP) and data science techniques, social media has recently been used for a variety of public health monitoring tasks in addition to pharmacovigilance [30]. These include monitoring the patterns of influenza [4], tracking tropical diseases like dengue fever [13], and analyzing disease outbreaks such as E. coli [10] and Ebola [27]. In behavioral medicine research, social media has been used to study users' lifestyle and analyzing the health-related choices they make. Researchers have used social media to study nutrition [34] and obesity patterns [23]. Applications also include analyzing alcohol [3], nicotine [31], and drug abuse [12]. There has been some research in timeline creation [20] and event extractions from timelines [36, 21, 8] for specific events. However, little effort has been invested in generating the summary of health-related data.

Only a handful of studies has attempted to predict pregnancy outcomes using quantitative data. Ines Banjari *et al.* [5] used clustering on a collection of questionnaire results accompanied by blood samples of 222 pregnant women who were in the first trimester. The authors performed hierarchical clustering considering three main features namely pre-pregnancy BMI, their age, and hemoglobin content. Via cluster analysis, the authors found that women with higher pre-pregnancy BMI and age have higher risks of complications during pregnancy. Laopaiboon *et al.* [18] studied the effect of maternal age and pregnancy outcome using health records of 308,149 singleton pregnant women. They used a multilevel, multivariate logistic regression with clustering technique to perform the study and found that 12.3% of these women had advanced maternal age (AMA) which varied across countries. Von Mandach *et al.* [37] studied 202 fetal disorders from Swiss ADR database using records

classified by regional pharmacovigilance centers as having ADRs. They performed a likelihood ratio and t-test and found that fetal disorders were closely associated with the ADRs of drugs they consumed. All these pregnancy-related studies have involved data sources from clinical records, reports, hospital patient data which often is expensive to obtain. Also, little information is available on lifestyle habits and drug usage after the patient's exit the medical facilities. Hence, social media and health forums are potentially attractive sources for extracting health information, drug usage patterns and their effects. Small samples of social media data have been used for performing pregnancy-related studies—such as the work by De Choudhury *et al.* [9], where 376 women were monitored to predict postpartum changes. Automatically collecting and processing large samples of social media data, however, presents significant challenges due to the lack of structure and use of informal language [32].

3. METHOD

Figure 1 gives a detailed illustration of our proposed system, which is broadly divided into three main steps: Data Collection and Classification, User Health Timeline Extraction, and Timeline Analysis. For data collection, we discuss how cohort timelines can be collected from the differing interfaces of Twitter and DailyStrength. For the last step of the analysis, we show how the timelines can be divided into pregnancy trimesters so that trimester-specific information, such as drug usage, can be further analyzed. Each of these steps is detailed in the following subsections.

3.1 Data Collection and Classification

Twitter and DailyStrength are the sources of our data for this study. We collected tweets originating from women announcing their pregnancy during a thirty-month time period, from January 2014 to September 2016. To identify our initial set of *potentially pregnant* women, we applied simple search expressions of the forms “*i’m * weeks/months pregnant*” and “*i am * weeks/months pregnant*”, with minor variations adding to 18 queries. DailyStrength has a very different structure compared to Twitter, and the website is divided into individual forums for specific cohorts. We obtained our data from five forums on DailyStrength (Pregnancy, Pregnancy After Loss Or Infertility, Pregnancy Teens, Stillbirth, and Miscarriage). We collected all the posts from all the users from these forums.

Due to the usage of relatively formal language and low noise, posts from DailyStrength are not subjected to pre-processing. In contrast, tweets contain an approximately equal share of useful information and noise in them. Hence, the tweets are pre-processed by removing URLs, user handles, emoticons, and stopwords. In the case of DailyStrength, because we collect posts from pregnancy-related forums we make the safe assumption that all users posting in the forums are currently pregnant or have been pregnant in the past. However, for Twitter data, manual inspection of a small sample of tweets revealed that approximately 35-40% of them were false positives (*i.e.*, posts that did not present personal admissions of pregnancy). Therefore, prior to collecting the timelines of the users making the announcement, we employ an automatic text classification technique to further filter out noisy tweets. We manually annotated 1200 randomly selected tweets (approximately 2% of all the collected tweets) mentioning pregnancy announcements into

isPreg (legitimate) and *notPreg* (not legitimate) classes.⁴ Some examples of pregnancy announcements and their annotations are shown in Table 1. The annotations were performed by two annotators, and the inter-annotator agreement (IAA) for was $\kappa = 0.79$, which is regarded as substantial agreement [17]. Disagreements were resolved by the third author of this paper, who reviewed the disagreement cases and performed the annotations independently. Among the 1200 tweets, 753 tweets were classified as *isPreg* and 447 were classified as *notPreg*.

Table 1: Sample tweets showing annotations for pregnancy announcements.

Tweet	Classification
<i>“I honestly still can’t believe I’m almost 5 months pregnant. Like wut.”</i>	<i>isPreg</i>
<i>“I can’t do this anymore. I work my ass off, and I’m eight months pregnant.”</i>	<i>isPreg</i>
<i>“I’m 18 weeks pregnant today and my 21st birthday is tomorrow. It’s a good day”</i>	<i>isPreg</i>
<i>“I’m 21 weeks, 5 days pregnant. Or, as I like to think of it: 128 days out from reclaiming my spot as my liquor store’s favorite customer. ”</i>	<i>isPreg</i>
<i>“It’s like just yesterday I was was at the doctor being told I was 14 weeks & 5 days pregnant .. now I’m 27 weeks & 3 days”</i>	<i>isPreg</i>
<i>“I hate how bloated I get when I’m on my period, like I look like I’m 3 months pregnant”</i>	<i>notPreg</i>
<i>“NOW. WAIT. ONE. MINUTE! I’m catching up on #LHHNY from last night, HOW did Mariah Lynn’s mufva go from 5 days pregnant to her 3rd trimester?”</i>	<i>notPreg</i>
<i>“I hate that I look at least 4 months pregnant every time I eat something wtf”</i>	<i>notPreg</i>
<i>“Girls will be two days pregnant already posting pictures talking bout “I’m getting big.””</i>	<i>notPreg</i>
<i>“My sister is five weeks and three days pregnant. I’m going to be an auntie oh my god””</i>	<i>notPreg</i>

Using the annotated data, we perform supervised classification of the tweets. We employ a variant of an existing social media text classification system [33],⁵ which were originally designed for adverse drug reaction detection. Past research on social media text classification suggests that an effective mechanism for classifying short Twitter posts is to generate large numbers of semantic features to balance the sparse word n-gram vectors. Therefore, for the classifier, we primarily remove adverse drug reaction specific features, keep the domain-independent features, and add some additional features. We briefly discuss some of the features in the following paragraphs.

⁴This the dataset will be made available with the final version of the paper.

⁵Source code for the classification system is available at: <https://bitbucket.org/asarker/adrbinaryclassifier>. Accessed on: 10/10/2016.

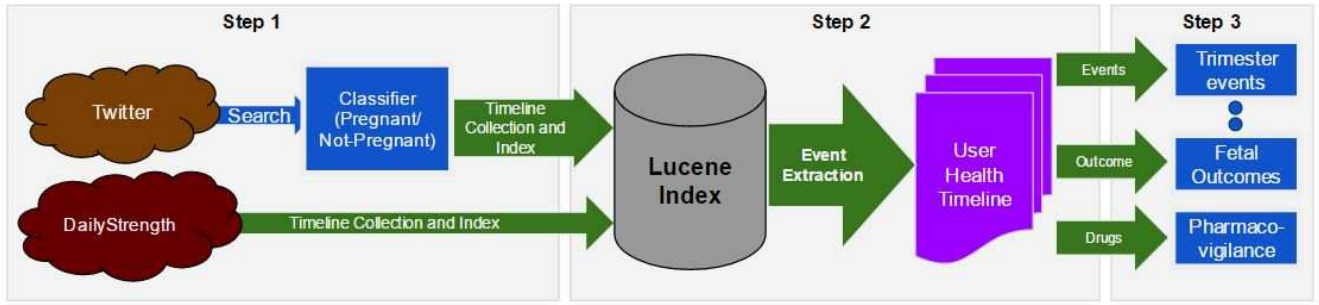


Figure 1: System architecture depicting our social media mining pipeline collecting and analyzing cohorts from social media.

N-grams and synsets

Word n-grams are the most common text classification features, consisting of sequences of contiguous n words in a text segment. We preprocess the texts by performing stemming and lowercasing, and use 1-, 2-, and 3-grams as features.

In addition to the words themselves, we use their synonyms in some cases to increase vocabulary coverage. For each adjective, noun or verb in a tweet, we use WordNet.⁶ to identify the synonyms of that term and add the synonymous terms as features.

Sentiment representing features

Our inspections of the announcements suggest that users generally express strong sentiments when making pregnancy announcements. So, we add features that express the sentiments of the users in various scales. We assign three sets of scores to sentences based on three different measures of sentiment. The first set of scores are derived from lists of positive and negative terms [16], the second set of scores are dependent on the prior polarities of terms present in a post [14], and the third set of scores are derived from a subjectivity lexicon that presents both polarity and subjectivity [38].

Word clusters

Recent research on social media based text classification suggests that using generalized representations of words, such as clusters of similar words, may improve performance [26]. In our work, we use the clusters generated by Owoputi *et al.* [29]. The authors generate the clusters by first learning vector representations of words [24] from over 56 million tweets, and then employing a Hidden Markov Model-based algorithm that partitions words into a base set of 1000 clusters, and induces a hierarchy among those 1000 clusters.⁷

To generate features from these clusters, for each tweet, we identify the cluster number of each token, and use all the cluster numbers associated with a tweet in a bag-of-words manner. Thus, every tweet is represented by a set of cluster numbers, with semantically similar tokens having the same cluster number.

Classification

Using these features, we trained Support Vector Machine (SVM) classifiers for the classification task. We used an RBF kernel, and we optimized the value of the cost parameter via 10-fold cross-validation over the 1200 annotated posts.⁸ We obtained the best results with cost=64.0, and we used this setting to classify all the identified tweets in our collection. Results of the classification are presented in the next section, including classification performance and the number posts (Table 2).

3.2 User Health Timeline Extraction

After the classification step, all the handles of the users classified to be legitimately pregnant are identified, and we attempt to collect their other posts using the Twitter API. We index all the posts into Apache Lucene⁹ for further analysis. Using the API, we collect all the user posts that are available from the past (*i.e.*, up to the limit allowed by Twitter) and sort them in chronological order, and we continue collecting tweets over time to monitor future health-related events.

For DailyStrength, however, since the forums we chose are all pregnancy-related, it is assumed that all users posting in these forums are/have been pregnant at some point during their membership to the website. The users can post across different forums on the website which can include interesting information such as drug intake admissions and adverse events. Hence, for each user posting a comment in one of the pregnancy related forums, we collect the user’s posts in all available forums to construct their timelines. Finally, we index each individual timeline into Apache Lucene with the following fields: userid, time, text, and trimester (if available) for further processing.

3.3 Timeline Analysis

Using the collected timelines, we attempt to explore if and how health-related events can be clustered into coarse-grained temporal windows. The duration of a pregnancy may be divided into three trimesters: first– week 1 through week 12, second– week 13 through week 27, and third– week 28 through birth. To successfully identify the trimester associated with a posted health-related event, information about

⁶<http://wordnet.princeton.edu/>. Accessed on: 01/05/2016.

⁷The clusters are publicly available at: <http://www.cs.cmu.edu/~ark/TweetNLP/>. Accessed on: 1/24/2017.

⁸We used the LibSVM implementation packaged with the python scikit-learn implementation: <http://scikit-learn.org/>. Accessed on: 10/11/2016.

⁹<https://lucene.apache.org/>. Accessed on: 11/23/2015.

the pregnancy start date is required. Via our manual inspections of the timelines, we discovered that pregnant mothers who announce their pregnancies over Twitter also often provide clues about the progress of the pregnancies. Consider the tweets below:

Oh well managed 8 out of 10 combat tracks, not bad at 28 weeks pregnant with the flu but still disappointing #frustrated

I'm officially 20 weeks pregnant & I've also never felt more sick in my life..)

The first tweet was posted during the third trimester and the second tweet was posted during the second trimester of pregnancy. Using this information, and the timestamp of the tweets, all the posts within a timeline can be grouped into the three trimesters. The key NLP challenge in this problem is to detect the statements regarding the progress of the pregnancies.

We use a combination of term and pattern matching algorithms to detect these trimester identifiers in each timeline. However, for Twitter, due to the 3200 tweet limitation enforced by the API, not all timelines that are extracted have all the tweets posted during the pregnancy time period. In our current algorithm, we first attempt to identify all tweets that mention the terms ‘pregnant’ and ‘pregnancy’ (seed word). Next, terms within a specified context window of the seed word are collected. Based on the empirical assessment, we settled for symmetric context window of size 6 terms. Within the context window, the algorithm then searches for *key* temporal terms such as ‘week’ and ‘month’, along with the presence of a number mention (e.g., six, 12, eighteen and so on). The number, along with the other mentioned terms are extracted and compared to the timestamp of the associated tweet to identify the trimester.

Following the organization of the timelines into trimesters, we assessed, in a preliminary fashion, if trimester-specific health events can be collected for further analysis. Depending on the intent of a study, the type of information that requires mining may vary, and detailed trimester-based health-related event analysis is outside the scope of this paper. Therefore, we simply focused on generating frequencies of the drugs that are mentioned at each trimester to make rough estimates about the drug usage patterns of the cohort at each phase. We perform a keyword search for each drug in Apache Lucene to obtain the drug mentions by users. Here, we make an assumption that all drug mentions are admissions of drug intake by the user. We query our Lucene index, and, for each drug, compute the number of users who have consumed it. The goal was to ascertain if a drug-usage information is available, rather than to perform a thorough analysis, which we leave as future work. Distributions of the drug mentions are presented in next section.

4. RESULTS AND DISCUSSIONS

The performance of our classifier was evaluated via 10-fold cross-validation, and the best results obtained are presented in Table 2. We compared the performance of the SVM to that of a Naïve Bayes baseline, which obtained an F-measure of 0.70 for the isPreg class. Figure 2 shows the ROC curves for each of the 10-folds of cross-validation, including the mean ROC for the positive class. The area under the mean ROC curve is 0.82. Running the SVM classifiers

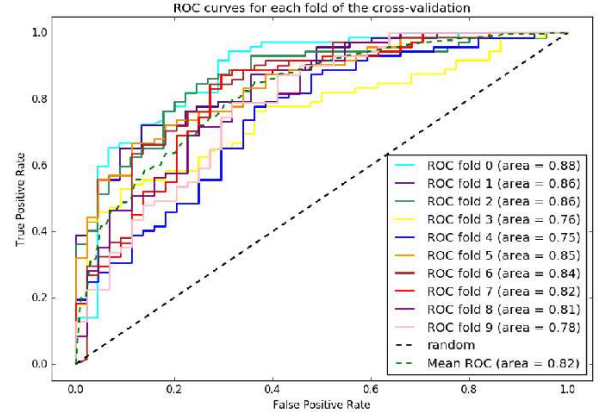


Figure 2: ROC curves for each fold of the 10-fold cross-validation, and the mean.

on our collected data resulted in the discovery of 34,895 legitimate pregnant women from a total of 53,820 users.

Table 2: Results from tweet classification for legitimate pregnancy announcements using SVM

Classification Result	Precision	Recall	F-measure
isPreg	0.83	0.79	0.81
notPreg	0.84	0.77	0.80

We applied our pregnancy trimester extraction algorithm on the 34,895 user timelines classified as legitimate pregnant users. Our algorithm detected pregnancy time-period for 15,523 (approximately 45%) users and was able to further categorize each tweet belonging to these timelines into one of the three trimesters. The remaining user handles were discarded from the analysis performed in the rest of this paper. We were able to collect over 30 Million tweets from these 15,523 users. Table 3 showcases a user timeline with the pregnancy trimester details and health-related tweets in each of the three trimesters.

We observe that the timeline contains health-related information such as drug intakes (in rows 1, 12, 20, 21, 22) and conditions/events (e.g., rows 1, 2, 4, etc.). We also notice that a large proportion of drug and condition mentions happen to be first-hand experiences. However, not all mentions of drugs are intakes (rows 24 and 25) and not all drug intakes are drug intakes by the user (rows 11 and 14). Similarly, not all conditions mentioned in the tweets are experienced by the user (rows 5 and 11). Mining drug intake and events are important in pharmacovigilance research for tracking ADRs. Hence, accurately distinguishing personal drug intake and events from mentions is an important NLP challenge that we intend to address in the future.

Trimester detection adds a very interesting NLP challenge. While some tweets are relatively easy to detect and were successfully processed by our rule-based algorithm, we found some that were missed or mis-classified. Consider the following Tweets, for example:

I b getting so much pressure next week is gone

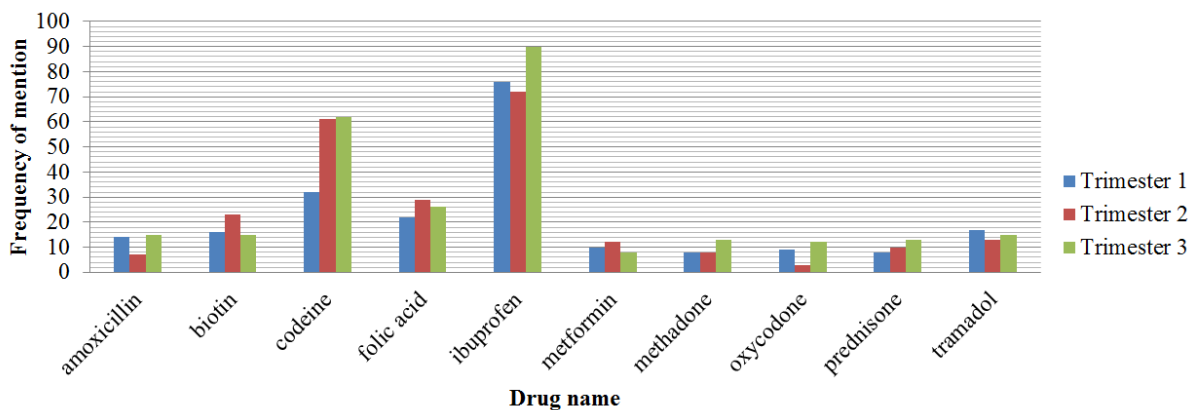


Figure 3: Distributions of top 10 drug mentions within the timelines across the three trimesters.

b my last week pregnant who want to make a bet lol

It is crazy to me that I am only 3 days past 13 weeks pregnant.

Our approach currently fails to detect the first tweet and mis-classifies the second tweet as first trimester instead of second. We leave the optimization of our detection algorithm as future work.

Figure 3 shows the distribution of popular drug mentions across the pregnancy trimesters for Twitter users. Even the most common drugs were mentioned by less than 0.5% of the users and the proportion of actual intakes may be lower. For instance, ibuprofen was one of the most common drugs mentioned in user timelines and it was mentioned by 76 unique users in their first trimester, 72 in their second, and 90 in their third. For a collection of more 15,000 user timelines, we find this proportion of mentions to be low for extensive analysis and hence we intend to expand our search terms and algorithms for cohort selection in the future.

For DailyStrength, our timeline collection approach retrieved a total of 257,531 posts from 11,435. In contrast to tweets, which are restricted to 140 characters, DailyStrength posts are longer. Thus, tracking the progress of pregnancies from their announcements to derive trimester information requires further NLP research. Discovering drug intake, however, is similar, and we find that common drug mentions within the user timelines in DailyStrength include drugs such as folic acid, aspirin, zolof and tylenol.

5. LIMITATIONS AND FUTURE WORK

We intend to build on this preliminary work in several key areas. Employing more sophisticated outcome detection techniques is an important future goal of this study. From the NLP perspective, our technique does not take into account tense (*e.g.*, past/present) and so the chronological order of posts may not represent the chronological ordering of events. Also, no mechanism is applied for gender detection among pregnancy announcement tweets, although our self-admission classifier does attempt to ensure that users included in the cohort are genuinely pregnant themselves.

Among other things, we intend to expand the drug list by including misspellings, spelling variations, phonetic variations and abbreviations of each drug. Similar to drug usage

pattern extraction, we could use a disease and disorder extraction method to classify mentions of diseases which would explain the reason why certain individuals consume a particular drug. As mentioned already in the paper, our trimester detection technique is currently not optimal, and we will improve it via the addition of more rules. Since the performance and effectiveness of the downstream applications and analyses depend heavily on the data collection and classification steps, our immediate focus will be to improve these. We will employ more queries to significantly increase the size of the cohort, and improve the performance of the classification step via the annotation of a much larger data set and the application of more sophisticated classification techniques. Ensembles of classifiers have been shown to perform particularly well for complex text classification tasks [2], and we will attempt to develop such systems with the view of maximizing recall while maintaining high precision.

6. CONCLUSION

In this paper, we presented the novel idea of collecting longitudinal health-related information about targeted cohorts from social media. We focused on the cohort of pregnant women in this study—a group that is not included in pre-market clinical trials. We presented a pipeline which includes three stages—identification of cohort, collection, and analysis. We discovered that large numbers of pregnant women can be identified with high-precision via a combination of rule-based and machine learning techniques. We discussed how health-related timelines can be gathered from two different social networks. Finally, we showed how temporal categorization of the timeline may be performed, and we verified that trimester-specific health-related information can be mined from the pre-processed timelines.

We discussed several limitations of our work, which will be addressed in future research. Crucially, while we focused solely on one cohort, our pipeline can be generalized for other population groups as well. This form of analysis may be particularly useful for population groups about whom data may not be available from other sources. In addition, social media may reveal information that people may not generally share via other means (*e.g.*, drug abuse/usage of illicit drugs). The results obtained by our current work are very promising and warrant future research.

7. REFERENCES

- [1] Cdc deaths in 2013. http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_02.pdf, 2013.
- [2] Automatic evidence quality prediction to support evidence-based decision making. *Artificial Intelligence in Medicine*, 64(2):89 – 103, 2015.
- [3] Y. Aphinyanaphongs, B. Ray, A. Statnikov, and P. Krebs. Text classification for automatic detection of alcohol use-related tweets. In *International Workshop on Issues and Challenges in Social Computing*, 2014.
- [4] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.
- [5] I. Banjari, D. Kenjerić, K. Šolić, and M. L. Mandić. Cluster analysis as a prediction tool for pregnancy outcomes. *Collegium Antropologicum*, 39(1), 2015.
- [6] J. Bian, U. Topaloglu, and F. Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM, 2012.
- [7] BusinessWire. Twenty six percent of online adults discuss health information online. <http://www.businesswire.com/news/home/20121120005872/en/Twenty-percent-online-adults-discuss-health-information>, 2012.
- [8] S. Choudhury and H. Alani. Personal life event detection from social media. 2014.
- [9] M. De Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM, 2013.
- [10] E. Diaz-Aviles and A. Stewart. Tracking twitter for epidemic intelligence: case study: Ehec/hus outbreak in germany, 2011. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 82–85. ACM, 2012.
- [11] L. B. Finer and S. K. Henshaw. Disparities in rates of unintended pregnancy in the united states, 1994 and 2001. *Perspectives on sexual and reproductive health*, pages 90–96, 2006.
- [12] N. Genes. Twitter discussions of nonmedical prescription drug use correlate with federal survey data. In *Medicine 2.0 Conference*. JMIR Publications Inc., Toronto, Canada, 2014.
- [13] J. Gomide, A. Veloso, W. Meira Jr, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd international web science conference*, page 3. ACM, 2011.
- [14] M. Guerini, L. Gatti, and M. Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *arXiv preprint arXiv:1309.5843*, 2013.
- [15] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91(6):1010–1021, 2012.
- [16] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
- [17] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- [18] M. Laopaiboon, P. Lumbiganon, N. Intarut, R. Mori, T. Ganchimeg, J. Vogel, J. Souza, and A. Gülmezoglu. Advanced maternal age and pregnancy outcomes: a multicountry assessment. *BJOG: An International Journal of Obstetrics & Gynaecology*, 121(s1):49–56, 2014.
- [19] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 117–125. Association for Computational Linguistics, 2010.
- [20] J. Li and C. Cardie. Timeline generation: Tracking individuals on twitter. In *Proceedings of the 23rd international conference on World wide web*, pages 643–652. ACM, 2014.
- [21] J. Li, A. Ritter, C. Cardie, and E. H. Hovy. Major life event extraction from twitter based on congratulations/condolences speech acts. In *EMNLP*, pages 1997–2007, 2014.
- [22] I. Lobo and K. Zhaurova. Birth defects: causes and statistics. *Nature Education*, 1(1):18, 2008.
- [23] Y. Mejova, H. Haddadi, A. Noulas, and I. Weber. #foodporn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 51–58. ACM, 2015.
- [24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [25] A. Nikfarjam and G. H. Gonzalez. Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1019. American Medical Informatics Association, 2011.
- [26] A. Nikfarjam, A. Sarker, K. O’Connor, R. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.
- [27] M. Odlum. How twitter can support early warning systems in ebola outbreak surveillance. In *143rd APHA Annual Meeting and Exposition (October 31-November 4, 2015)*. APHA, 2015.
- [28] A. C. O’Higgins. A survey of the use of social media by women for pregnancy. In *Medicine 2.0 Conference*. JMIR Publications Inc., Toronto, Canada, 2013.
- [29] O. Owoputi, B. O’Connor, C. Dyer, K. Gimpel, and N. Schneider. Part-of-speech tagging for twitter: Word clusters and other advances. *School of Computer Science*, 2012.
- [30] M. J. Paul, A. Sarker, J. S. Brownstein, A. Nikfarjam, M. Scotch, K. L. Smith, and G. Gonzalez. Social mining for public health monitoring and surveillance.

- Pacific Symposium of Biocomputing*, 21:468–479, 2016.
- [31] K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson. Identifying health-related topics on twitter. In *Social computing, behavioral-cultural modeling and prediction*, pages 18–25. Springer, 2011.
 - [32] A. Sarker, R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya, and G. Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics*, 54:202–212, 2015.
 - [33] A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207, 2015.
 - [34] S. Sharma and M. De Choudhury. Detecting and characterizing nutritional information of food and ingestion content in instagram. *Proc. WWW Companion*, 2015.
 - [35] S. Sinclair, M. Cunningham, J. Messenheimer, J. Weil, J. Cragan, R. Lowensohn, M. Yerby, and P. Tennis. Advantages and problems with pregnancy registries: observations and surprises throughout the life of the international lamotrigine pregnancy registry. *Pharmacoepidemiology and drug safety*, 23(8):779–786, 2014.
 - [36] M. Wen, Z. Zheng, H. Jang, G. Xiang, and C. P. Rosé. Extracting events with informal temporal references in personal histories in online communities. In *ACL (2)*, pages 836–842, 2013.
 - [37] C. Wettach, J. Thomann, C. Lambrigger-Steiner, T. Buclin, J. Desmeules, and U. von Mandach. Pharmacovigilance in pregnancy: adverse drug reactions associated with fetal disorders. *Journal of perinatal medicine*, 41(3):301–307, 2013.
 - [38] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

Table 3: Excerpts from a Twitter user’s health timeline in reverse chronological order

No	Tweet	Trimester	Category
1	<i>“God bless Zantac for helping me not want to throw up from the insane heartburn I’ve been having. #pregnancyproblems #37weeks”</i>	third	drug, condition
2	<i>“Ugh. Awful dreams; was obviously clenching my teeth all night - woke up with sore jaw; a headache. #ouch #tired #iwanttogobacktobed”</i>	third	condition
3	<i>“Went to the chiropractor today and my lower back actually feels worse right now. #sore”</i>	third	condition
4	<i>“If I don’t drink Powerade before bed, I get awful leg cramps. If I do, I get awful heartburn. #cantwin #PregnancyProblems”</i>	third	condition
5	<i>“Betting I’ll be back at Dr. AnonX’s again real soon. Anonymized has a 101.8 degree fever. Why do my kids get sick so much?! #mommyproblems #sickkids”</i>	third	condition
6	<i>“W/ the leg pain from the INSANE cramp I had this AM; the pelvic pain from how baby’s laying, the waddle is strong today. #pregnancyproblems”</i>	third	condition
7	<i>“Fell asleep around 10:30. Just woke up sweating and uncomfortable. It’s too hot in here to sleep now. #PregnancyProblems”</i>	third	condition
8	<i>“Lunch was absolutely AMAZING (Tuscan chicken; artichoke soup from @anonymized) but now I have the worst heartburn. #PregnancyProblems”</i>	third	condition
9	<i>“Baby’s moving around like crazy; the pain in my side/back is FINALLY letting up. RELIEF! #pregnancyproblems”</i>	third	condition
10	<i>“My pelvis hurts so freaking bad right now and I still have 14 weeks to go til my due date. #PregnancyProblems”</i>	second	condition
11	<i>“At the doctor with Anon again. Fever was 102 this AM, gave her Tylenol, then up to 104 after her nap. Her only complaint - being cold.”</i>	second	drug, condition
12	<i>“@anonymized Taking multivitamin gummy in AM, calcium + D in afternoon, prenatal at bedtime....just like I always have.”</i>	second	drug
13	<i>“Well this is a new one - my Vitamin D level is actually too HIGH. Now OB wants me to see bariatric doc/nutritionist again.”</i>	second	drug
14	<i>“It’ll be another sleepless night checking on Anon periodically. 104.8 degree fever earlier, Motrin brought it down to 100.1”</i>	second	drug
15	<i>“So help me if I’m getting ANOTHER cold I’m gonna be pissed. Scratchy sore throat and runny nose all of a sudden.”</i>	second	condition
16	<i>“Slim chance it’s the start of appendicitis. If I get a fever, nausea/vomiting/diarrhea,; pain is worse, I need to come back ASAP.”</i>	second	condition
17	<i>“Weak gag reflex + coughing + snot = disaster waiting to happen. I just want to go home, crawl in bed, and sleep until this cold is gone.”</i>	second	condition
18	<i>“Not bring able to take anything for this stupid cold is awful. So miserable. #stuffedup #cantbreathe #cough”</i>	second	condition
19	<i>“It’s amazing how a headache, raging hormones; lack of sleep make you want to stuff a pillow in the face of a snoring husband. #shutup”</i>	second	condition
20	<i>“For the second time in a row, Tylenol PM has left me wide awake at 3AM after passing out for a whopping 5 hours. #needmoresleep”</i>	second	drug
21	<i>“I love the Olympics! Too bad I just took some Tylenol PM that’s starting to kick in so I won’t be watching much longer tonight.”</i>	second	drug
22	<i>“AnonT stayed home from work today on daddy duty so I just popped some Tylenol PM and I’m sleeping this cold away. #goodnight”</i>	first	drug
23	<i>“My throat is dry; sore from breathing through my mouth but breathing through my nose isn’t possible right now. #snot #sick”</i>	first	condition
24	<i>“Woke up to anonymized crying around 5:30; a skull-knocker headache. And guess what...no Tylenol except for PM stuff. #crap”</i>	first	drug
25	<i>“Been sleeping awful so AnonT said to take Tylenol PM; get a good nights sleep. Come up to bed; discover we’re out of Tylenol PM. #gofigure”</i>	first	drug, condition
26	<i>“Finally start to feel drowsy so I try to sleep, but then I get all twitchy and can’t lay still. #insomnia #sleepproblems”</i>	first	condition
27	<i>“Pseudo gallbladder attacks in the middle of the night are awesome, said no one ever. #pain”</i>	first	condition
28	<i>“Just turned my head and something popped in my neck the wrong way. Big ouch. Need to hit up the chiropractor on my way home. #hurt”</i>	first	condition
29	<i>“Gotta pack up the kids; head to MQT. Woke up with a skull knocker headache so me thinks it’s time for a back; neck cracking. #chiropractor”</i>	first	condition